Abstract
Potential gender inequalities in Wikipedia articles
along different dimensions: notability, topical focus, linguistic bias, structural
properties, and meta-data presentation.

(i) women in Wikipedia are more notable than men, outcome of a subtle glass ceiling effect;
(ii) family-, gender-, and relationship-related topics are more present in biographies about
women;
(iii) linguistic bias since abstract terms tend to be used to
describe positive aspects in the biographies of men and negative aspects in the
biographies of women
(iv) there are structural differences in meta-data and hyperlinks, which affect
information-seeking activities.

1. Introduction

Objectives: we aim to address the following research questions:
(i) Are men and women who are depicted in Wikipedia equally notable
(ii) Are any topical aspects overrepresented in articles about men or women?
(iii) Does linguistic bias manifest in Wikipedia?
(iv) Do articles about men and women have similar structural properties, i.e., similar meta-data, and network properties in the hyperlink network?

Our results show that:
• Women in Wikipedia are on average slightly more notable than their male counterparts. Furthermore, the gap between the number of men and women is larger for 'local heroes' (people who are only depicted in few language editions) than for 'superstars' (people who are present in almost all language editions)

- Gender-, family-, and relationship-related topics are more dominant in the stand-alone overviews of biographies about women in the English Wikipedia.
Abstract terms tend to be used to describe positive aspects in biographies of men, and negative aspects in biographies of women.
- structural differences in terms of meta-data and hyperlinks, which have consequences for information-seeking activities

Our analysis of the global notability of men and women in Wikipedia reveals that women are slightly more notable than men using internal and external proxy measures for notability

Further, the men-to-women ratio is higher than expected for local heroes (i.e. people who only show up in language edition) and lower for superstars. These findings suggest the existence of a subtle glass-ceiling effect that makes it more difficult for women to be included in Wikipedia than for men

At least three plausible explanations exist that describe why the glass-ceiling effect may

be present in Wikipedia: () the narrow diversity of editors may foster the glass-ceiling effect since it is well known that individuals generally favor people from their in-group over people from their out-group [,]; () men are potentially more likely to create an article about themselves since previous research suggests that men are on average more self-absorbed than women []; () the external materials on which Wikipedia editors rely may introduce this bias, since the life of women or certain ethnic minorities may be less well documented and less visible on the Web.

2 Data and methods

To study gender bias in Wikipedia, we consider the following data sources:
. The DBpedia dataset
. Inferred gender for Wikipedia biographies by

We split this dataset in Pre- and Post-. The Pre- sample contains all people born before , while the Post- sample consists of people born in or after .

The number of biographies, proportion of biographies about women and the biography overlap with the English edition are depicted. One can see that the fraction of women on average around 17% and the average overlap with English is 97%

Approach

To assess the extent to which gender bias manifests in Wikipedia, we compare Wikipedia articles about men and women along the following dimensions:

. Global notability of people according to external and internal proxy measures.
. Topical focus and linguistic bias of biography articles.
. Structural properties of articles, including meta-data and network-theoretic position of people in the Wikipedia article link network.

Global notability

the glass-ceiling effect refers to the situation in which women cannot reach higher positions because an ‘invisible barrier’ (namely, gender bias) prevents them from doing so.
We hypothesize that if the entry point of Wikipedia functions as a glass ceiling, fewer women will be included in Wikipedia, but those women will be more notable than their male counterparts on average. Especially if we compare the number of male and female ‘local heroes’ (people with low levels of notability, without worldwide fame), we expect to see a larger gender gap (i.e., fewer women than men) than for worldwide ‘superstars,’ because fewer female ‘local heroes’ will be able to make it into Wikipedia.

The number of Wikipedia language editions that contain an article about a person is used as an internal proxy measure for that person’s global notability. The idea is that people who only show up in a few language editions are less relevant from a global perspective than those who show up in more language editions.

Concretely, we use the following external and internal proxy measures:

- Number of language editions: The number of Wikipedia language editions that contain an article about a person is used as an internal proxy measure for that person's global notability.

To explore whether the number of editions is influenced by gender, we fit a negative binomial (NB) regression model. The number of editions in which a person is depicted is used as dependent variable, while gender is used as independent variable.

- Google search volume: serve as an external proxy for the public interest toward a person

2.2.2 Topical and linguistic bias

women were relegated to matters of 'social and purely feminine affairs' and as subjects, women were often little more than addenda to male biographies (e.g., Marie Curie as the wife of Pierre Curie)

for members of our in-group, we tend to describe positive actions and attributes using more abstract language, and their undesirable behaviors and attributes more concretely. In other words, we generalize their success but not their failures. Note that verbs are usually used to make more concrete statements (e.g., 'he failed in this play'), while adjectives are often used in abstract statement (e.g., 'he is a bad actor'). Conversely, when an out-group individual does or is something desirable, we tend to describe them with more concrete language (we do not generalize their success), whereas their undesirable attributes are encoded more abstractly (we generalize them)

Topical bias: To unveil topical biases in Wikipedia content, we analyze the following three topics that could be over-represented in articles about women according to what is suggested by Thomas's observations in Britannica and the Finkbeiner test:

• The gender topic contains words that emphasize that someone is a man or woman (i.e., man, women, mr, mrs, lady, gentleman) as well as sexual identity (e.g., gay, lesbian).

• The relationship topic consists of words about romantic relationships (e.g., married, divorced, couple, husband, wife).

• The family topic aggregates words about family relations (e.g., kids, children, mother, grandmother).

Linguistic bias: To measure linguistic bias, we use a lexicon-based approach and syntactic annotations to detect abstract and subjective language as proposed by Otterbacher []. The level of abstraction of language can be detected through the syntactic class of terms, where adjectives are the most abstract class, as for example comparing ‘is violent’ with ‘hurt the victims’

2.2.3 Structural properties

Structural properties impact how visible and reachable articles about notable men and women are, since users and algorithms rely on this information when navigating Wikipedia or when assessing the relevance of content within a certain context. For instance, search result rankings are often informed by centrality measures such as PageRank.

Meta-data: we compare the relative proportions of attribute presence between genders using chi-square tests, considering the male proportion as baseline, and discuss which differences go beyond what can be explained by professional areas

Hyperlink network
compute the PageRank of articles about people. PageRank is a widely used measure of network centrality [,]. To explore potential asymmetries in network centrality, we sort the list of biographies according to their PageRank values in descending order. We estimate the fraction of biographies that are about women at different ranks k. , we compare our empirical results with those obtained from baseline graphs that are constructed as follows: Random, Degree Sequence and Small World method

Result

3.1 Inequalities in global notability thresholds

Let us first test our hypothesis that the Wikipedia entry point functions as a glass ceiling, making it more difficult for women to be included. If this is the case, women who made it into Wikipedia should be more notable than men.

Figure shows that since , the gap between men and women is indeed larger for people with low or medium level of global notability than for the ‘global superstars,’ compared to a baseline. If we focus on strictly local heroes (people who only appear in one language edition), the men to women ratio is larger than expected by chance.

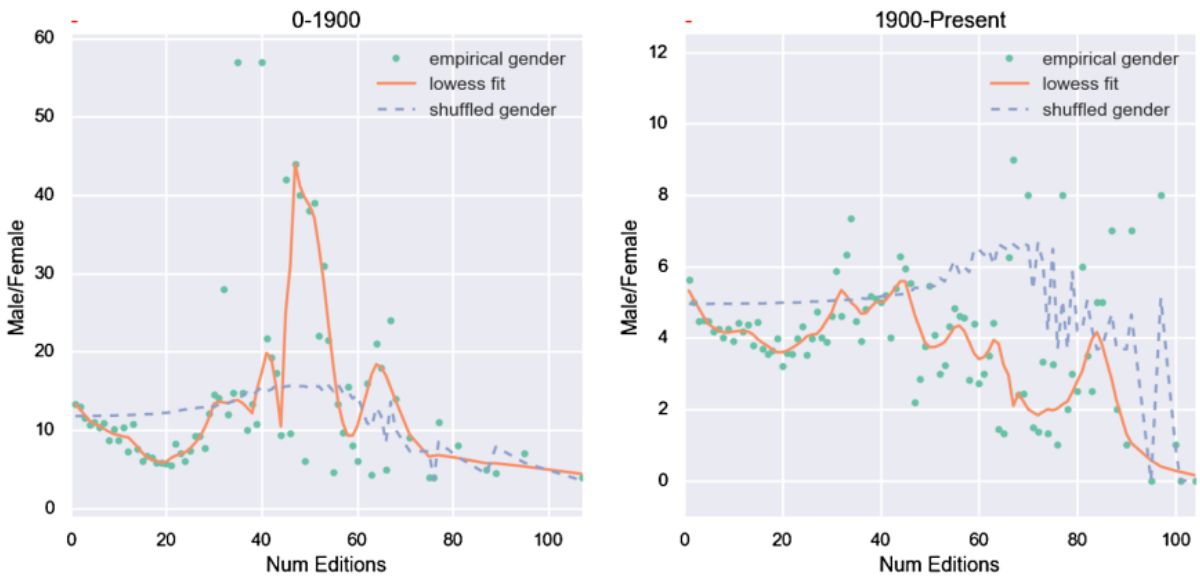

Figure 1 Men-women ratio. Ratio of men to women included in N language editions before 1900 (left) and since 1900 (right), as a function of N. The gender gap since 1900 is larger for people with low or medium global notability than for the global superstars. The empirical ratio is smoothed with locally weighted regression (solid lines) and compared with a baseline obtained by random shuffling of genders (dashed lines)

A possible explanation for the high men-to-women ratio for local heroes is that the entry barrier into Wikipedia is higher for women than for men. Note that people can also create articles about themselves in Wikipedia; men are on average more self-absorbed than women [], and thus may be more likely to create articles about themselves. Another possible explanation is that more information may be available online about less notable men than about less notable women. Since Wikipedia editors rely on secondary information sources, their decisions also reflect the biases that exist in other media.

If we only look at people born since , we see
that women are % more notable than men, while limiting our dataset to people born before indicates that women are % less notable than men. For people in Wikipedia born before , being a female decreases the chances of notability, as one would predict based on the historical exclusion of women []. Conversely, for people in Wikipedia born since , being female increases the chances of notability. Due to the noted relation between being historic and global notability (see Figure), we cannot claim a glass-ceiling effect for inclusion in Wikipedia of women born prior to .
The model further indicates that the decade when a person was born is negatively associated with notability; the more historic a person is, the more
notable they are from a global perspective. This is expected: people from older centuries

appear on Wikipedia because their ideas and actions have transcended time (through secondary sources). Conversely, people of recent fame can be notable in terms of availability of secondary sources, but not necessarily because their ideas will remain valuable in time

When we consider
people born since we find that Wikipedia developed a ‘recency bias’; people in this group are slightly more notable if they were born more recently
younger people may benefit from the greater availability of digital information about them or generated by them, making them more likely to be recognized by Wikipedia editors.

3.1.2 Google search trends
women in Wikipedia are slightly more of interest to the world according to Google’s relative search volume statistics.
results suggest that the gender of a person that made it into Wikipedia is significantly related to the number of regions and months in which this person is of interest,
we cannot exclude other confounders. For example, women included in Wikipedia tend to be born in recent years (see Figure) and people born in recent years may have received more attention on Google between and

3.2 Topical and linguistic asymmetries
3.2.1 Topical bias

Beside professional and topical areas, words in the gender, relationship,
and family categories are more dominant in articles about women born before 1900.
Gender-specific
differences are much less pronounced in articles about people born since 1900

Women tend to have more words related to family, gender and
relationships than men.

3.2.2 Linguistic Bias
adjectives are almost % more likely to be used
to describe positive aspects of men’s biographies, while .% more likely to describe negative aspects in women’s biographies

3.3 Structural inequalities
3.3.1 Meta-data
Attributes activeYearsEndDate, activeYearsStartYear, careerStation,
numberOfMatches, position, team, and years are more frequently used to describe men. All of these attributes are related to sports, therefore the differences can be explained by the prominence of men in sports-related
Attributes deathDate and deathYear are more frequently used for men born before . A possible explanation is that the life of women was less well documented than

Attribute birthName is more frequently used for women in recent times. Its value refer mostly to the original name of artists, and women have considerable presence in this class []. A likely explanation is that married women change their surnames to those of their husbands in some cultures.

The spouse attribute is more frequently used for women in recent times. This attribute indicates whether the portrayed person was married or not, and with whom

women biographies have more links to other women articles than one would expect by chance. A possible explanation for this asymmetry
stems from the reported interests of female editors, who frequently edit biographies about women in Wikipedia
empirically observed structure of the hyperlink network puts women (especially women born since 1900) at a
disadvantage when it comes to ranking algorithms
one must conclude that there exists a bias in the generation of
links by Wikipedia editors, favoring articles about men

4. Discussions

Our analysis of the global notability of men and women in Wikipedia reveals that women are slightly more notable than men using internal and external proxy measures for notability

Further, the men-to-women ratio is higher than expected for local heroes (i.e. people who only show up in language edition) and lower for superstars. These findings suggest the existence of a subtle glass-ceiling effect that makes it more difficult for women to be included in Wikipedia than for men

At least three plausible explanations exist that describe why the glass-ceiling effect may be present in Wikipedia: () the narrow diversity of editors may foster the glass-ceiling effect since it is well known that individuals generally favor people from their in-group over people from their out-group [,]; () men are potentially more likely to create an article about themselves since previous research suggests that men are on average more self-absorbed than women []; () the external materials on which Wikipedia editors rely may introduce this bias, since the life of women or certain ethnic minorities may be less well documented and less visible on the Web.

One way to mitigate the glass-ceiling effect is by relaxing notability guidelines for women, in order to include women who are locally notable

The topical and linguistic asymmetries that we found highlight that editors need to pay attention to the ways women are portrayed in Wikipedia

Even though the structural inequalities that we found suggest that editors (especially

those who edit articles about women) do a great job in interlinking articles about women, the visibility of women is still lower than expected when link-based ranking algorithms such as PageRank are applied.

Wikipedia should provide tools to help editors, for instance, by considering already existing manuals of gender-neutral language [], or by indicating missing links between articles. For example, if an article about a woman links to the article about her husband, the husband should also link back

Our empirical results are limited to the English Wikipedia, which is biased towards western cultures []. However, in previous work [] we found that similar structural, topical and coverage biases exist across six different language editions. We leave a more detailed exploration of gender bias across all language editions for future work. Our methods can be applied in other contexts given an ad-hoc manual coding of associated keywords to each gender.

(i) we presented a computational method for assessing gender bias in Wikipedia along multiple dimensions and (ii) we applied this method to the English Wikipedia and shared empirical insights on observed gender inequalities. The methods presented in this work can be used to assess, monitor and evaluate these issues in Wikipedia on an ongoing basis. We translate our findings into some potential actions for the Wikipedia editor community to reduce gender biases in the future. We hope our work will contribute to increased awareness about gender biases online, and about the different ways these biases can manifest themselves. We propose that Wikipedia may wish to consider revising its guidelines, both to account for the low visibility of women and to encourage a less biased use of language.